

Diccionarios y computadoras

Innovaciones aplicadas a la lexicografía

Hecsil Coello

Licenciada en letras por la Universidad Católica Andrés Bello. Acaba de egresar en el Máster en Lexicografía en la Escuela de Lexicografía Hispánica de la Real Academia Española.

Tradicionalmente la Lexicografía se ha dedicado a elaborar diccionarios, delimitando el léxico con el propósito de definir los significados asociados a una palabra o pieza léxica. Para ello los lexicógrafos han manejado un considerable número de textos –inicialmente obras literarias–, de los que seleccionaban tanto las palabras como las citas de los literatos que certificaban el buen uso de la lengua –la norma–, para luego anotar la información en papeletas que serían ordenadas y almacenadas en ingentes ficheros, consultados luego por el lexicógrafo al definir. Así fue elaborado en 1726 el *Diccionario de Autoridades* y el *Diccionario de la lengua castellana* en 1780, y sus posteriores ediciones.

Desde el siglo XVIII la metodología para delimitar y describir el léxico fue básicamente la misma, pero desde el siglo XIX y con mucha mayor fuerza desde mediados del siglo XX, la Lexicografía española se desprende del canon establecido por la Real Academia desde *Autoridades*, el *Diccionario de la lengua castellana* y posteriormente el *DRAE*, apareciendo la llamada Lexicografía no académica que introdujo innovaciones paulatinas en la metodología para elaborar diccionarios. El cambio metodológico más destacado de esta nueva Lexicografía ocurre en las últimas décadas del siglo XX al pasar del enfoque normativo, según el cual priva el buen uso de las palabras, al enfoque descriptivo, cuyo interés reside en mostrar el uso de la lengua. Esto permitió a los diccionarios editados en ese momento un enfoque más paralelo, más usual de la lengua española. Lo que se concreta en estas publicaciones fundamentalmente en el de-

sapego de usos que resultan ya anacrónicos, en la introducción de nuevas acepciones para una palabra o bien en la introducción de nuevas entradas, haciendo del diccionario una obra más cercana y provechosa. Pueden citarse obras fundamentales como el *Diccionario de uso del español* de María Moliner y el *Diccionario del español actual* de Manuel Seco.

Especialmente con obras como la de Seco la Lexicografía no académica comienza a elaborar diccionarios basados en corpus lexicográficos, entendiendo éstos como copiosos conjuntos de textos principalmente escritos, pero provenientes ahora de diversas fuentes de información, literarias o no, delimitados cronológicamente y distribuidos según un porcentaje que favorece un tipo de texto sobre otro, dependiendo del tipo de diccionario a elaborar. Este cambio de paradigma supuso una primera innovación relacionada principalmente con la descripción de las piezas léxicas de la lengua; no obstante, la técnica para documentar el léxico seguía siendo similar a los siglos anteriores, se buscaba en las fuentes de información y se papeletizaban los datos. La innovación residió básicamente en pasar de la recolección manual a la mecánica y luego a la digital, –puesto que comenzaron a escanearse los textos, y las papeletas o fichas se elaboraban en las computadoras–, aunque el almacenamiento continuaba haciéndose en ficheros.

Si bien la Lexicografía no académica se había propuesto describir el léxico, el corpus lexicográfico suponía, tal y como se había diseñado, un filtro para la eficiente explotación de la información contenida

en éste. El papeletizar los datos entrañaba el riesgo de pasar por alto factores verdaderamente relevantes en el estudio de una pieza léxica. Piénsese en que la selección de unidades estaba aún sesgada por el trabajo individual y la intuición del lexicógrafo, no teniéndose así garantía de que serían abarcados todos los significados y posibilidades lingüísticas que pudieran ocurrir en una pieza léxica, según la variación del contexto en que ésta apareciera. Así para lograr una real descripción del léxico eran necesarias aún herramientas que permitieran obtener información sistematizada de manera eficaz y eficiente.

Lingüística computacional y creación de corpus electrónicos

En las últimas décadas del siglo XX, debido a los avances e innovaciones de la informática y de la lingüística computacional, se crean herramientas versátiles que han modificado la metodología lexicográfica, en cuanto a la compilación y almacenamiento de datos léxicos. Se trata de cuantiosas bases de datos en las que el lexicógrafo dispone de millones de unidades junto a sus contextos, ordenadas en lista, a las que puede tener acceso rápida y eficazmente, sin elaborar ni consultar fichas. Estas bases de datos no han de confundirse con el mero escaneo y almacenamiento electrónico de miles de textos, se trata de corpus electrónicos creados por lingüistas computacionales, que han seleccionado, digitalizado, almacenado y distribuido en un porcentaje cada vez más equitativo todo tipo de textos –tanto escritos como orales–, de todas las variedades dialectales del español, a los que han agregado una amplia información lingüística. El propósito de este tipo de corpus es obtener en una consulta toda la información relevante de una pieza léxica, puesto que están diseñados para recuperar listados de palabras junto a su contexto y hacer búsquedas realmente minuciosas de índole gramatical e incluso semántica.

Para lograr la recuperación de datos y hacer explícita la información lingüística contenida en los textos, es necesario anotar o codificar cada unidad que aparece en éste, mediante un complejo proceso que permite que cada pieza léxica sea clasificada –dependiendo de los requerimientos del diseño del corpus– según su fonética, el tipo de palabra (sustantivo, adjetivo, verbo, adverbio, preposición) o la función

que cumple en la oración (sujeto, complemento directo, indirecto, circunstancial); siguiendo pues esta clasificación los mismos niveles del sistema de la lengua. Nótese que al enfoque descriptivo se une la tecnología para explotar hondamente el contenido léxico de los textos de un corpus.

Con una búsqueda en un corpus electrónico se obtiene información estructural muy diversa, puede observarse por ejemplo si una misma pieza léxica tiene cambios en la categoría gramatical, como el caso de adjetivos que también funcionan como sustantivos o que según su significado pueden tener uso solamente en femenino y no en masculino, o bien su significado se restringe a una variante dialectal del español, es decir, que el significado de la pieza léxica depende del país en que se diga. Obsérvense, por ejemplo la pieza léxica “gordo” definida en el *Diccionario Clave*, editado por la editorial SM:

gordo, da

adj. 1 Grueso, abultado o voluminoso: *Este jersey es muy gordo y abriga mucho.* 2 Grave, importante o fuera de lo corriente: *Tengo un problema bastante gordo.* **adj./s.** 3 Referido a una persona o a un animal, que tiene muchas carnes o grasas: *Si no estuvieras tan gordo no te fatigarías tanto.* **s.m.** 5 *col.* Grasa de la carne animal: *Dejó el gordo del filete en el plato.* **s.f.** 6 En zonas del español meridional, tortilla de maíz más gruesa y más pequeña que la común: *Voy a preparar gorditas de frijoles para merendar.* 7 armarse la gorda *col.* Organizarse un alboroto: *Cuando vino el jefe y vio que el trabajo no estaba hecho, se armó la gorda.* caer gordo *col.* Referido a una persona, resultar antipática: *No sé por qué, pero me cae gordo.* ni gorda *col.* Nada o casi nada: *Sin gafas no veo ni gorda.*

ETIMOLOGÍA: Del latín *gurdus* (embotado).

MORFOLOGÍA: 1. En la acepción 3, la RAE solo lo registra como adjetivo. 2. En la acepción 6, se usa mucho el diminutivo *gordita*.

Puede verse resaltado en letra negrita en las acepciones 1, 3, 5 y 6 el cambio en la categoría gramatical y el contraste que hace esta obra con el *DRAE* en el apartado morfología, en el que se indica que el diccionario académico únicamente registra una categoría gramatical, mientras que el *Clave* registra dos e incluso especifica el uso femenino de “gorda” que se corresponde además con una variante geográfica del español. Puede verse también en

la acepción 5 la marca “col”, relativa al uso coloquial de esta acepción de la palabra, lo cual se repite en las expresiones subrayadas “armarse la gorda”, “caer gordo”. He ahí la utilidad de los corpus electrónicos, el que el lexicógrafo puede evaluar los listados de ejemplos en que aparece una misma pieza léxica y pueda ir registrando acepciones, usos y valores de una misma palabra, apegados al uso de la lengua, logrando capturar gran información que enriquece y engrosa los artículos de los diccionarios.

La Lexicografía española cuenta con corpus lingüísticos que han sido usados para la elaboración de diccionarios, tal es el caso del corpus “Cumbre”, usado en la elaboración del *Gran diccionario de Uso del Español Actual*, publicado en 2001, y los elaborados por la Real Academia Española: el “Corpus de Referencia del español actual, CREA”, que contiene textos producidos entre 1975 y 2008, usado en la elaboración del *Diccionario del estudiante* y el “Corpus Diacrónico del español, CORDE”, que contiene textos desde los inicios del español hasta 1975. CORDE actualmente es usado –junto a otras bases de datos– en la elaboración del *Nuevo diccionario histórico de la lengua española*. Tanto CREA (<http://corpus.rae.es/creanet.html>), como CORDE (<http://corpus.rae.es/cordenet.html>) tienen acceso libre desde la página electrónica de la Academia; otro corpus de libre acceso es el de Mark Davies, el “Corpus del español” (<http://www.corpusdelespanol.org/x.asp>), usado para diversos estudios lingüísticos.

Corpus lingüístico, rica fuente de datos para el lexicógrafo

Con esta herramienta electrónica el lexicógrafo consulta y recibe datos totalmente insospechados, inesperados e incluso desconocidos, he ahí la ventaja de trabajar con un corpus que no ha elaborado el propio lexicógrafo. Recuperar datos que contienen información lingüística sobre los cuales éste estudiará el comportamiento léxico y tomará decisiones. Por ello la metodología de delimitación y descripción del diccionario basada en el análisis de corpus supone una ventaja considerable y una garantía con respecto a la forma tradicional de recolectar datos.

El hecho de que la pieza léxica aparezca junto a su contexto y el corpus lo or-



© Cortesía de César Fernández-Arias

dene en listas de ejemplos, permite al lexicógrafo evaluar minuciosamente el comportamiento léxico de una unidad, posibilitándole tareas realmente provechosas. Tómese por ejemplo hacer una búsqueda según la terminación de una palabra, esto permite desmitificar la tradicional no inclusión en el diccionarios de ciertas palabras derivadas, como el caso de los adverbios terminados en -mente de los que se ha dicho que su significado es fácilmente deducible de la palabra de la cual deriva, sin embargo resulta curioso que al consultar en los corpus se encuentren significados relacionados con valores cuantitativos y no modales. Esto quiere decir que hay significados de adverbios como “sumamente”, que no se relacionan con la categoría tradicional de adverbio de modo, puesto que no significaría “de manera suma” sino “en sumo grado”, ese “grado” da a este adverbio un valor de cantidad. Lo cual lleva al lexicógrafo a reflexionar sobre este tipo de palabras, estudiarlas en su contexto y rediseñar su tratamiento en el diccionario, ya sea incluyéndolas, o bien agregando otras acepciones, o bien reestructurando las pocas definiciones que de ellas hay en los diccionarios. He ahí otra ventaja de la delimitación y descripción del léxico a partir del análisis de corpus y el acceso a al uso real de las palabras.

Corpus lingüístico y nuevo paradigma metodológico

La elaboración de diccionarios a partir del análisis de corpus supone pues un cambio de paradigma en el estudio de las unidades léxicas y por ende un cambio en la metodología de elaboración de diccionarios. En primer lugar, el análisis de corpus plantea al lexicógrafo la posibilidad de recuperar una gran cantidad de datos que le permiten evaluar el contexto léxico de una palabra, por ende sus posibles significados y cambios de significado, su estructura gramatical y cambios de esta estructura, la frecuencia de uso de una palabra, lo que incide inmediatamente sobre el uso o desuso de la misma, o bien la preferencia de uso sobre una palabra y no sobre otra, es una rica fuente de investigación para los neologismos y palabras especializadas, fructíferos en la lengua a partir de los avances científicos, tecnológicos y de toda área de especialidad.

Elaborar diccionarios apoyados en la lingüística de corpus responde a una metodología de investigación novedosa, que nace de la visión descriptiva del léxico pero que ha evolucionado a una manera innovadora de establecer la norma en el sistema de la lengua. Se trata de crear corpus o consultar los ya existentes y estudiar el repertorio léxico que contendrá el diccionario para extraer del corpus los significados reales y usuales de una pieza léxica y a partir de los resultados establecer la norma, siguiendo un criterio de frecuencia de uso y cambio de significado y no el tradicional criterio normativo afianzado por una autoridad. Es por ello menester para la lexicografía de hoy hacer diccionarios a partir de ejemplos y el com-

promiso del lexicográfico consiste en trabajar con el análisis de corpus.

Vale la pena inquietarse ante la necesaria creación de corpus lingüísticos representativos del español, o bien representativos de una variedad dialectal, es decir, de corpus electrónicos particulares del español de Venezuela, de Colombia, por ejemplo, o bien de corpus particulares de zonas dialectales como el español del Caribe, cuyo acceso sea necesariamente libre. Esta innovación metodológica aporta a la lexicografía grandes posibilidades a la hora de elaborar diccionario más abarcadores –como el diccionario histórico– balanceados, en otros formatos e incluso otro tipo de obras como los diccionarios fraseológicos, que permiten consultar combinaciones de palabras del tipo “armarse la gorda”; diccionarios de régimen en los que se consulta qué preposición rige a un sustantivo o verbo, como en los casos de “pensar en”, “ir a”; “ir por”; los de valencia que permiten saber cuáles y cuántos complementos tiene una determinada palabra; o bien los diccionarios de colocaciones y combinatorias léxicas, obras en las que no hay definiciones sino agrupaciones de palabras que dan cuenta de las posibilidades combinatorias de unas palabras con otras, según la relación gramatical y de ideas que hay entre ellas, es decir, se encuentra en la entrada del verbo “resolver” una lista de palabras que se combinan con éste, por ejemplo, los sustantivos “conflicto” y “contratiempo”, son obras muy útiles para crear textos. La Lexicografía española cuenta con dos obras fundamentales de este tipo, se trata de *Redes, diccionario combinatorio del español contemporáneo* y el *Diccionario combinatorio práctico del español contemporáneo* de Ignacio Bosque. ◀▶